Here is a clearer description of the Bonferroni procedure for multiple comparisons than what I rushed in class.

If there are m hypothesis tests and we want a procedure for which the probability of rejecting one or more hypotheses falsely (when the corresponding null hypotheses are true) to be less than a prespecified significance level α, we can accomplish this by testing each hypothesis separately and rejecting only if the p-value is less than α/m. So, for example, if there are ten possible pairwise comparisons, if we set $\alpha = 0.05$ and if all the population means were equal, there is no more than a 5% chance that any single p-value from one of the ten pair-wise t-tests would be smaller than $0.05/10 = 0.005$.

The Bonferroni procedure is completely general — it applies to any set of hypothesis tests, not just pairwise comparisons between population means — but is also quite conservative — the actual probability of at least one Type I Error (at least one incorrect rejection of a true null hypothesis) may have probability quite a bit smaller than α. For very large numbers of comparisons it is essentially worthless. But it is reasonable for a small number of comparisons.

Here is the data from Exercise 11.24.

Treatment	A	B	C	D	E
Mean	4.37	4.76	3.70	5.41	5.38
n	9	9	9	9	9

The pooled estimate of the common standard deviation is $s = \sqrt{0.2246} = 0.474$. There are a total of $45 - 5 = 40$ degrees of freedom within samples. The standard error for a difference in sample means using the pooled estimate is

$$\mathrm{SE}(\bar{y}_i - \bar{y}_j) = s\sqrt{\frac{1}{n_i} + \frac{1}{n_j}}$$

There are two equivalent ways to apply the Bonferroni method to test all possible pairwise comparisons if we want a family-wise error probability (the probability of one or more rejected null assuming all nulls to be true) to be no more than $\alpha = 0.05$. We could carry out each possible t-test (with the pooled SE) and compute a two-sided p-value for each by comparison to a t distribution with 40 df, rejecting only those tests where $p < 0.05/10 = 0.005$, or we could compute each possible pairwise 99.5% confidence interval and reject those hypotheses corresponding to confidence intervals that do not contain 0.

Here is the confidence interval approach. The margin of error is particularly easy because all of the sample sizes are equal. Here is the computation of the common margin of error using R.

```
> sp <- sqrt(0.2246)
> sp

[1] 0.4739198

> se <- sp * sqrt(1/9 + 1/9)
> se

[1] 0.2234079

> tcrit <- qt(0.9975, 40)
> me <- tcrit * se
> me

[1] 0.6637833
```

A 99.5% confidence interval goes between the 0.0025 and 0.9975 quantiles. A confidence interval for $\mu_i - \mu_j$ will not contain 0 if and only if the difference in sample means $\bar{y}_i - \bar{y}_j$ is larger (in absolute value) than the margin of error.

Here is a way in R to compute all the pairwise differences and to indicate which are significant. The Bonferroni procedure calls any pairwise differences greater than 0.664 significant.

```
> ybars <- c(4.37, 4.76, 3.7, 5.41, 5.38)
> diffs <- matrix(0, 5, 5)
> for (i in 1:5) {
+     for (j in 1:5) {
+         diffs[i, j] <- ybars[i] - ybars[j]
```

```
+     }
+ }
> rownames(diffs) <- LETTERS[1:5]
> colnames(diffs) <- LETTERS[1:5]
> diffs

      A     B    C     D     E
A  0.00 -0.39 0.67 -1.04 -1.01
B  0.39  0.00 1.06 -0.65 -0.62
C -0.67 -1.06 0.00 -1.71 -1.68
D  1.04  0.65 1.71  0.00  0.03
E  1.01  0.62 1.68 -0.03  0.00

> sig <- diffs
> sig[diffs > me] <- "+"
> sig[diffs < -me] <- "-"
> sig[(-me < diffs) & (diffs < me)] <- ""
> sig

  A   B   C   D   E
A ""  ""  "+" "-" "-"
B ""  ""  "+" ""  ""
C "-" "-" ""  "-" "-"
D "+" ""  "+" ""  ""
E "+" ""  "+" ""  ""
```

The results here are slightly different than the Newman-Keuls method in class. If we ranked the treatments in order from smallest to largest, C A B E D, there would be a line under B through D and an overlapping line from A to B. B cannot be distinguished (significantly) from A, D, or E, but A is significantly less than both E and D. C is significantly less than all the others.

For the t-test approach, we would compute each t statistic and then compare these to the 0.9975 quantile of the t distribution with 40 degrees of freedom. Here is a way to do this using R.

```
> ts <- diffs/se
> ts

          A         B        C          D          E
A  0.000000 -1.745685 2.998998 -4.6551612 -4.5208777
B  1.745685  0.000000 4.744683 -2.9094757 -2.7751922
C -2.998998 -4.744683 0.000000 -7.6541592 -7.5198757
D  4.655161  2.909476 7.654159  0.0000000  0.1342835
E  4.520878  2.775192 7.519876 -0.1342835  0.0000000

> sig2 <- ts
> sig2[ts > tcrit] <- "+"
> sig2[ts < -tcrit] <- "-"
> sig2[(-tcrit < ts) & (ts < tcrit)] <- ""
> sig2

  A   B   C   D   E
A ""  ""  "+" "-" "-"
B ""  ""  "+" ""  ""
C "-" "-" ""  "-" "-"
D "+" ""  "+" ""  ""
E "+" ""  "+" ""  ""
```

It may also be interesting to examine the individual p-values.

```
> pvalues <- 2 * pt(-abs(ts), 40)
> pvalues

             A            B            C            D            E
A 1.000000e+00 8.854346e-02 4.642551e-03 3.528049e-05 5.372888e-05
B 8.854346e-02 1.000000e+00 2.661254e-05 5.886584e-03 8.347788e-03
C 4.642551e-03 2.661254e-05 1.000000e+00 2.340353e-09 3.577681e-09
D 3.528049e-05 5.886584e-03 2.340353e-09 1.000000e+00 8.938518e-01
E 5.372888e-05 8.347788e-03 3.577681e-09 8.938518e-01 1.000000e+00

> round(pvalues, 4)

       A      B      C      D      E
A 1.0000 0.0885 0.0046 0.0000 0.0001
B 0.0885 1.0000 0.0000 0.0059 0.0083
C 0.0046 0.0000 1.0000 0.0000 0.0000
D 0.0000 0.0059 0.0000 1.0000 0.8939
E 0.0001 0.0083 0.0000 0.8939 1.0000

> signif(pvalues, 2)

        A       B       C       D       E
A 1.0e+00 8.9e-02 4.6e-03 3.5e-05 5.4e-05
B 8.9e-02 1.0e+00 2.7e-05 5.9e-03 8.3e-03
C 4.6e-03 2.7e-05 1.0e+00 2.3e-09 3.6e-09
D 3.5e-05 5.9e-03 2.3e-09 1.0e+00 8.9e-01
E 5.4e-05 8.3e-03 3.6e-09 8.9e-01 1.0e+00
```

Using Bonferroni, only those p-values less than 0.005 are deemed to be significant.